

Report Heart disease risk study

Comparing classification models for predicting the risk of a heart attack.





KANAV MALIK

Contents

[**INTRODUCTION** 3](#_Toc57913885)

[Importance 3](#_Toc57913886)

[Description 3](#_Toc57913887)

[**METHODOLOGY** 4](#_Toc57913888)

[Part 1 4](#_Toc57913889)

[Part 2 4](#_Toc57913890)

[**DATA ANALYSIS** 5](#_Toc57913891)

[Part 1 5](#_Toc57913892)

[Part 2 5](#_Toc57913893)

[Summary of results 6](#_Toc57913894)

[**CONCLUSION** 7](#_Toc57913895)

[**APPENDIX** 8](#_Toc57913896)

## INTRODUCTION

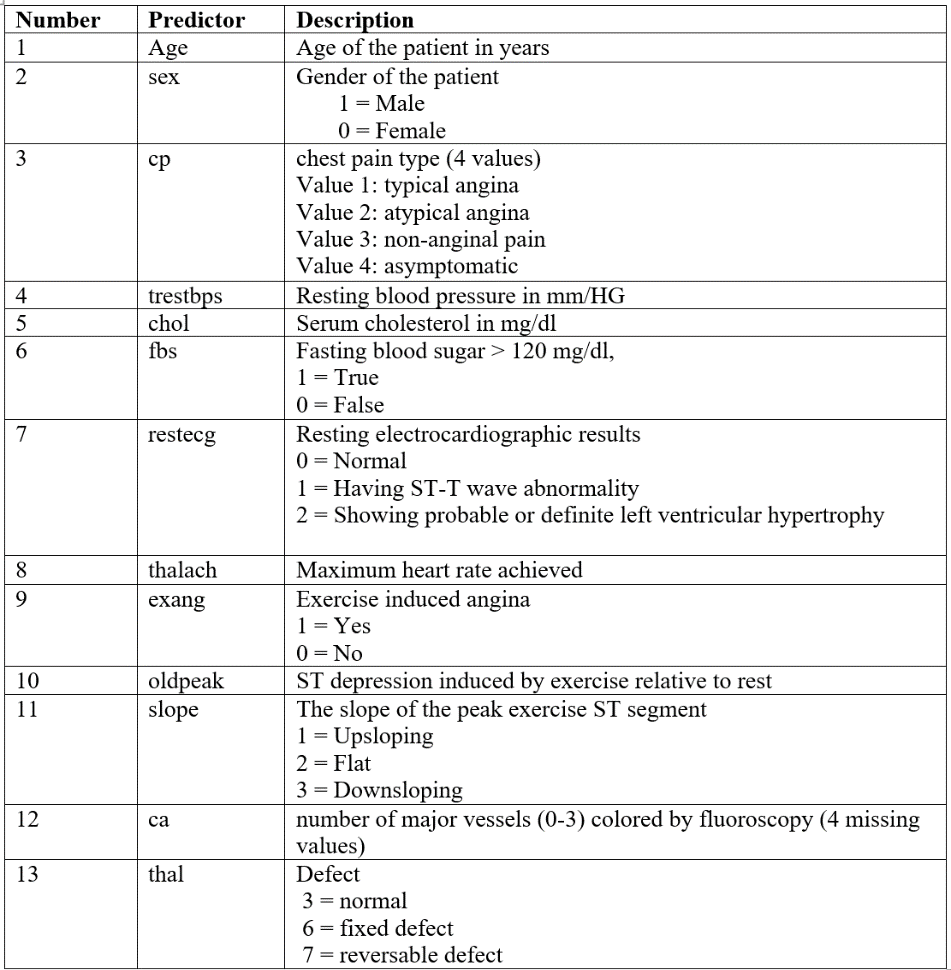
# Importance

In the United States one person dies every 36 seconds making heart disease one of the leading causes of death for men and women1. This disease has an estimated cost of 219 billion dollars each year in lost productivity, medicines, and health care costs1. The signs of a heart are silent leading the person to be unaware of it or the factors that lead up to it until it is to late. There are several different factors that can increase a person’s risk of developing heart disease. Which is why applying statistical classification methods to these potential risk factors can be beneficial in helping doctors predict a patient’s risk and lead to early detection of the disease. This is why we chose the Cleveland Heart Data set to conduct our analysis2.

# Description

TheCleveland dataset consists of 76 diagnostics and histories of patients in relation to the presence or absence of heart disease. Of these 76 a subset of 13 diagnostic factors (Table 1) were chosen for the goal of building and comparing different classification models for predicting the outcome of heart disease.

**Table 1.** Summary and Description of 13 diagnostic factors in relation to the dependent variable of heart disease status. (Target 0 = Absence, 1 = Present)



## METHODOLOGY

# Part 1

***Data Exploration / Logistic Regression***

Data exploration will aide in summarizing the size, initial patterns, and accuracy of the heart dataset. By using correlation plot, bar graphs, and boxplots will help visualize possible relationships and interactions of the13 diagnostic factors with the target variable.

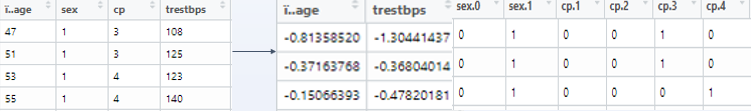
# Part 2

***Clustering***

K-means unsupervised clustering would aid us in checking the optimal number of clusters for the observations.

***K-Nearest Neighbors / Support Vector Machines***

These 2 non-parametric distance-based approaches will be applied to the 80% sampled training dataset. The confusion matrix will be computed to compare the actual vs estimated results.

These methods work on scaled interval variables, so the interval variables have been scaled and the categorical variables have been converted to dummy variables. The results should make sense after the data has been modified. This is how the data looked before and after modification 

***Naïve Bayes Classifier using Kernel Density Estimate***

It is a parametric based simple technique for constructing classifier which is based on a strong assumption that the value of a particular feature is independent of the value of any other feature , given the class variable. An advantage of this method is that it requires only a small training dataset to estimate the parameters necessary for classification.

## DATA ANALYSIS

# Part 1

***Data Exploration / Logistic Regression***

***Clustering***

Although, we are working on a supervised dataset, it is interesting to see the optimal number of clusters for our dataset.

The unsupervised K-means clustering gives us the optimal number of clusters that are appropriate, given our dataset. We find that the optimal number of clusters given our dataset are 2.

# Part 2

***K-Nearest Neighbors***

10-fold cross validation on the training dataset is done to get the optimal K as 10. The complete set of results can be seen in Appendix 2.

***Support Vector Machines***

10-fold cross validation on the training dataset is done to get the optimal parameters for “linear”, ”polynomial”, ”radial” density as 10. The complete set of results can be seen in Appendix 2.

***Naïve Bayes Classifier***

The complete set of results can be seen in Appendix 2.

# Summary of results

Here is a summary of the results –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Optimal Parameters** | **Train MSE** | **Test MSE** | **Test Sensitivity** | **Test Specificity** |
| Logistic Regression | age, sex, cp, testbps, col,  thalalach, slope, ca, thal |  | 85.9% |  |  |
| KNN | K = 10 | 86.9% | 83.3% | 88.6% | 76.0% |
| SVM - Linear Kernel | cost = 0.1 , gamma = 0.05 | 86.5% | 88.3% | 91.4% | 84.0% |
| SVM - Polynomial Kernel | cost = 0.1 ,degree = 1 , gamma = 0.05 | 86.5% | 90.0% | 97.1% | 80.0% |
| SVM - Radial Kernel | cost = 0.85 , gamma = 0.05 | 94.5% | 91.7% | 97.1% | 84.0% |
| Naive Bayes - KDE | prior prob(0) =0.525,  prior prob(1) =0.475(comes from data) | 84.4% | 85.0% | 85.7% | 84.0% |

## CONCLUSION

The SVM radial kernel performs the best based on the Train and Test MSE’s. It seems to be a good model for predicting the true positives and true negatives and scores better on all metrics of accuracy, sensitivity, specificity, false negatives. However, if we compare the models based on Specificity or False Positives rate, then logistic regression performs way better than any other model(91% specificity).

If the task if of predicting whether a person has the risk of heart attack, then we would like to have a low False positive rate i.e. low chances of incorrectly classifying patients at risk as not being at a risk of heart attack. So if the purpose is of having a low False positive rate, then logistic regression is the best model to use and if the purpose is increasing the sensitivity(low false negative rate), then SVM radial kernel is the best model to use. Finally, if the task is inference for research, then logistic regression is the only parametric method to use out of these methods.

## APPENDIX

1. CDC. Heart Disease. In. (

2. Dua, D.a.G., Casey. (2017). {UCI} Machine Learning Repositor. In. (University of California, Irvine, School of Information and Computer Sciences.

Appendix



